

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 05-242065
(43)Date of publication of application : 21. 09. 1993

(51)Int. CI. G06F 15/18
G06F 15/16
G06F 15/80
G06G 7/60

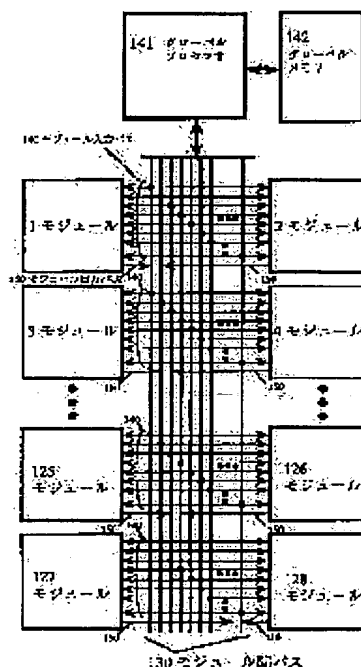
(21)Application number : 04-042830 (71)Applicant : HITACHI LTD
(22)Date of filing : 28.02.1992 (72)Inventor : ASAI MITSUO
SHIBATA KATSUNARI
SATO YUJI
YAMADA MINORU
SAKAGUCHI TAKAHIRO
HASHIMOTO MASA

(54) INFORMATION PROCESSOR AND ITS SYSTEM

(57)Abstract:

PURPOSE: To provide a very compact information processing system which permits fast neural net learning and operation, permits operation based upon instruction sets different from each other for modules and also permits large-scale ultra parallel calculation.

CONSTITUTION: One module (modules 1-128) is constituted with a function block for arithmetic operation (capable of fast inner product operation) and a control part for controlling it. In a module, operation is made in SIMD method. Multiple modules can be connected for communication. Multiple modules operate in MIMD method which includes different instruction set. In addition, within one module, the sections which are hardly affected with defects of a neuron processor and memory are accumulated on an integrated circuit board. As far the control part which is easily affected with the detects, the



integrated circuit board of the control part is mounted on the integrated circuit board using, for example, silicon-on-silicon technology.

LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number].

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998, 2003 Japan Patent Office

(19) 日本国特許庁(JP)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平5-242065

(43) 公開日 平成5年(1993)9月21日

| | | | | |
|----------------------------|-------|------------------|-----|--------|
| (51) Int. Cl. ⁵ | 識別記号 | 庁内整理番号 | F I | 技術表示箇所 |
| G 0 6 F | 15/18 | 8945-5 L | | |
| | 15/16 | 3 9 0 Z 9190-5 L | | |
| | 15/80 | 9190-5 L | | |
| G 0 6 G | 7/60 | | | |

審査請求 未請求 請求項の数30

(全21頁)

(21) 出願番号 特願平4-42830

(22) 出願日 平成4年(1992)2月28日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 浅井 光男

東京都国分寺市東恋ヶ窪1丁目280番地 株

式会社日立製作所中央研究所内

(72) 発明者 柴田 克成

東京都国分寺市東恋ヶ窪1丁目280番地 株

式会社日立製作所中央研究所内

(72) 発明者 佐藤 裕二

東京都国分寺市東恋ヶ窪1丁目280番地 株

式会社日立製作所中央研究所内

(74) 代理人 弁理士 小川 勝男

最終頁に続く

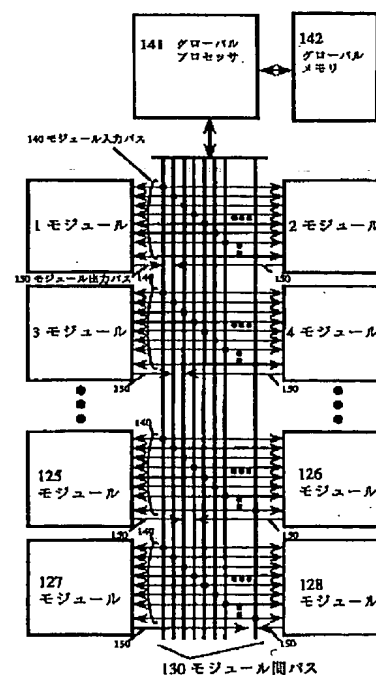
(54) 【発明の名称】 情報処理装置及びシステム

(57) 【要約】

【構成】 演算を行う機能ブロック(内積演算を高速に行えるもの)とこれを制御する制御部を1つのモジュール(1~128モジュール)とする。モジュール内ではSIMD方式により動作を行う。さらに複数個のモジュールを接続し通信を可能とする。複数個のモジュールは、異なる命令セットであるMIMD方式により動作を行う。また、1つのモジュール内では、ニューロンプロセッサ及びメモリなどの欠陥に対して強い部分は集積回路基板上に集積し、欠陥に対して弱い制御部は、例えばシリコン・オン・シリコン技術により、その集積回路基板上に制御部の集積回路基板を搭載する。

【効果】 ニューラルネットの学習及び動作を高速に行うことができる。更にモジュールごとに異なる命令セットにより動作させることが可能である。また大規模の超並列計算を行う情報処理システムを非常にコンパクトに提供できる。

図1



【特許請求の範囲】

【請求項1】内積演算を行なう複数のニューロンプロセッサを第1の集積回路基板上に集積し、上記複数のニューロンプロセッサを制御する制御プロセッサを第2の集積回路基板上に集積し、上記第1の集積回路基板上に上記第2の集積回路基板を搭載して単一のモジュール構成としたことを特徴とする情報処理装置。

【請求項2】請求項1の情報処理装置において、上記第1及び第2の集積回路基板をシリコンで構成し、シリコン・オン・シリコンにより上記第2の集積回路基板を搭載したことを特徴とする情報処理装置。

【請求項3】請求項1の情報処理装置において、上記複数のニューロンプロセッサは、夫々演算器と重み値を保持するメモリを備え、上記複数のニューロンプロセッサを相互に接続するデータベースを備えたことを特徴とする情報処理装置。

【請求項4】請求項1の情報処理装置において、上記複数のニューロンプロセッサに対する動作命令を格納するエリアを有するワーキングメモリを上記第1の集積回路基板上に集積したことを特徴とする情報処理装置。

【請求項5】請求項1の情報処理装置において、アナログ信号を入力しデジタル信号に変換するA/D変換器を第3の集積回路基板上に集積し、上記第1の集積回路基板上に上記第3の集積回路基板を搭載して単一のモジュール構成としたことを特徴とする情報処理装置。

【請求項6】重み値を保持するメモリ、その重み値を用いて内積演算を行う演算器、及びその内積演算の結果を保持する手段を備えた複数のニューロンプロセッサと、上記複数のニューロンプロセッサを相互に接続するデータベースと、少なくとも上記複数のニューロンプロセッサに対する動作命令を格納するワーキングメモリと、上記動作命令を上記複数のニューロンプロセッサに出力するための命令バスとを少なくとも第1の集積回路基板上に集積し、上記複数のニューロンプロセッサを制御する制御プロセッサを第2の集積回路基板上に集積し、上記第1の集積回路基板上に上記第2の集積回路基板を搭載して単一のモジュール構成としたことを特徴とする情報処理装置。

【請求項7】請求項6の情報処理装置において、上記制御プロセッサは、上記ワーキングメモリと、上記複数のニューロンプロセッサの内積演算の結果を保持する手段のメモリとを同一のメモリ空間としてアクセスすることとを特徴とする情報処理装置。

【請求項8】演算器と情報を保持するメモリを備えた複数の機能ブロックと、上記複数の機能ブロックを相互に接続し、データを通信する手段を第1の集積回路基板上に集積し、

上記複数の機能ブロックに対する動作命令の発生を制御し、上記複数の機能ブロックとメモリ空間を共有するスカラープロセッサを第2の集積回路基板上に集積し、上記第1の集積回路基板上に上記第2の集積回路基板を搭載して単一のモジュール構成としたことを特徴とする情報処理装置。

【請求項9】相互に接続された複数の第1の機能ブロックと、上記複数の第1の機能ブロックに対して同一の動作命令を発生する第2の機能ブロックとを備えた情報処理装置を、複数個相互に接続したことを特徴とする情報処理システム。

【請求項10】請求項9の情報処理システムにおいて、上記情報処理装置は、演算器と情報を保持するメモリを夫々備えた上記複数の第1の機能ブロックと、上記複数の第1の機能ブロックを相互に接続するデータベースと、少なくとも上記複数の第1の機能ブロックに対する動作命令セットを格納し、上記複数の第1の機能ブロックを制御する上記第2の機能ブロックと、上記動作命令セットを上記複数の第1の機能ブロックに出力するための命令バスとを備え、上記複数の第1の機能ブロックは、上記動作命令セットに従って動作することを特徴とする情報処理システム。

【請求項11】請求項9の情報処理システムにおいて、上記情報処理装置の上記複数の第1の機能ブロックを単一の集積回路基板上に集積したことを特徴とする情報処理システム。

【請求項12】請求項9の情報処理システムにおいて、各々の上記情報処理装置は、同一または異なるニューラルネットモデルの計算を行うことを特徴とする情報処理システム。

【請求項13】請求項12の情報処理システムにおいて、上記ニューラルネットモデルとして、相互結合型ニューラルネットワークを用いることを特徴とする情報処理システム。

【請求項14】データの演算を行う複数の機能ブロックを備え、上記複数の機能ブロックは単一の命令セットに従って単一命令複数データ流方式(SIMD)で動作する情報処理装置を複数個設け、上記複数個の情報処理装置間を非同期に通信し、複数命令複数データ流方式(MIMD)で動作させることを特徴とする情報処理システム。

【請求項15】請求項14の情報処理システムにおいて、各々の上記情報処理装置は、当該情報処理装置以外の他の情報処理装置より順次入力される信号を順次記憶する記憶手段と、上記複数の機能ブロックと上記順次記憶する記憶手段とを相互に接続するデータベースと、上記データベース上に出力された所定の機能ブロックの出力を他の情報処理装置へ順次出力する手段とを備え、上記情報処理装置間是非同期に相互通信することを特徴とする情報処理システム。

【請求項16】請求項15の情報処理システムにおいて、各々の上記情報処理装置は、単一命令複数データ流方式（SIMD）で動作する上記複数の機能ブロックに対し、上記単一命令セットの発生を制御するスカラープロセッサを備えたことを特徴とする情報処理システム。

【請求項17】請求項14の情報処理システムにおいて、上記情報処理装置の上記複数の機能ブロックを単一の集積回路基板上に集積したことを特徴とする情報処理システム。

【請求項18】内積演算を行う複数のニューロンプロセッサを備え、上記複数のニューロンプロセッサは単一の命令セットに従って単一命令複数データ流方式（SIMD）で動作する情報処理装置を複数個設け、上記複数の情報処理装置間を非同期に通信し、複数命令複数データ流方式（MIMD）で動作させることを特徴とする情報処理システム。

【請求項19】請求項18の情報処理システムにおいて、上記複数の上記情報処理装置のうち少なくとも2個の情報処理装置は、同一のニューラルネットモデルを計算を行うことを特徴とする情報処理システム。

【請求項20】請求項19の情報処理システムにおいて、上記ニューラルネットモデルとして、相互結合型ニューラルネットワークを用いることを特徴とする情報処理システム。

【請求項21】内積演算を行う複数のニューロンプロセッサと、上記複数のニューロンプロセッサに対する単一の命令セットの発生を制御するスカラープロセッサとを備え、上記スカラープロセッサは上記複数のニューロンプロセッサのメモリ空間を含んでアクセスし、単一命令複数データ流方式（SIMD）で動作する情報処理装置を複数個設け、上記複数の情報処理装置間を非同期に通信し、複数命令複数データ流方式（MIMD）で動作させることを特徴とする情報処理システム。

【請求項22】内積演算を行う複数のニューロンプロセッサと、上記複数のニューロンプロセッサに対する単一の命令セットの発生を制御する第1のスカラープロセッサとを備え、単一命令複数データ流方式（SIMD）で動作する情報処理装置を複数個設け、上記複数の情報処理装置に対する複数の命令セットの発生を制御する第2のスカラープロセッサを備え、上記複数の情報処理装置間を非同期に通信し、複数命令複数データ流方式（MIMD）で動作させ、上記第2のスカラープロセッサは、上記第1のスカラープロセッサ、及び上記複数のニューロンプロセッサのメモリ空間を含んでアクセスすることを特徴とする情報処理システム。

【請求項23】情報を保持するメモリ、その情報を用いて演算を行う演算器、及びその演算の結果を保持する手段を備えた複数の機能ブロックと、上記複数の機能ブ

ックに対する単一の命令セットの発生を制御する第1のスカラープロセッサと、上記単一の命令セットを格納するエリアを有し、上記第1のスカラープロセッサのワーキング用のメモリとを備え、単一命令複数データ流方式（SIMD）で動作する情報処理装置を複数個設け、上記複数の情報処理装置に対する複数の命令セットの発生を制御する第2のスカラープロセッサと、上記複数の命令セットを格納するエリアを有し、上記第2のスカラープロセッサのワーキング用のメモリとを備え、上記複数の情報処理装置間を非同期に通信し、複数命令複数データ流方式（MIMD）で動作させ、上記第2のスカラープロセッサは、上記第2のスカラープロセッサのワーキング用のメモリと、上記複数の機能ブロックの演算の結果を保持する手段のメモリと、上記第1のスカラープロセッサのワーキング用のメモリを同一のメモリ空間としてアクセスすることを特徴とする情報処理システム。

【請求項24】請求項23の情報処理システムにおいて、上記情報処理装置の上記複数の機能ブロックを、単一の集積回路基板上に集積したことを特徴とする情報処理システム。

【請求項25】請求項24の情報処理システムにおいて、各々の上記情報処理装置は、同一または異なるニューラルネットモデルの計算を行うことを特徴とする情報処理システム。

【請求項26】請求項25の情報処理システムにおいて、上記ニューラルネットモデルとして、相互結合型ニューラルネットワークを用いることを特徴とする情報処理システム。

【請求項27】演算器と情報を保持するメモリを備えた複数の機能ブロックと、上記複数の機能ブロックを相互に接続するデータバスと、少なくとも上記複数の機能ブロックに対する動作命令を格納するワーキングメモリと、上記動作命令を上記複数の機能ブロックに出力するための命令バスとを少なくとも第1の集積回路基板上に集積し、上記複数の機能ブロックに対する動作命令の発生を制御するスカラープロセッサを第2の集積回路基板上に集積し、外部のI/O装置とのインタフェースを行うI/Oプロセッサを第3の集積回路基板上に集積し、上記第1の集積回路基板上に上記第2及び第3の集積回路基板を搭載して単一のモジュール構成としたことを特徴とするワークステーション。

【請求項28】請求項27のワークステーションにおいて、上記スカラープロセッサは、上記ワーキングメモリと、上記複数の機能ブロックの情報を保持するメモリとを同一のメモリ空間としてアクセスすることを特徴とす

るワークステーション。

【請求項29】視覚情報を処理する複数の機能ブロックを備え、上記複数の機能ブロックは単一の命令セットに従って単一命令複数データ流方式(SIMD)で動作する情報処理装置と、

聴覚情報を処理する複数の機能ブロックを備え、上記複数の機能ブロックは単一の命令セットに従って単一命令複数データ流方式(SIMD)で動作する情報処理装置とを設け、

上記情報処理装置間を非同期に通信し、複数命令複数データ流方式(MIMD)で動作させることを特徴とするロボット制御システム。

【請求項30】動画像の1フレームの画像情報を処理する複数の機能ブロックを備え、上記複数の機能ブロックは単一の命令セットに従って単一命令複数データ流方式(SIMD)で動作する情報処理装置を複数個設け、上記複数個の情報処理装置間を非同期に通信し、複数命令複数データ流方式(MIMD)で動作させ、複数フレーム間の動画像処理を行うことを特徴とする動画像処理システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、情報処理装置及びシステムに係り、ニューロコンピュータを始め、並列計算機、物理シミュレーションシステム、ワークステーション、ロボットの制御システム等、非常に幅広く利用することができるものである。

【0002】

【従来の技術】従来の並列計算機の実現方法で粒度の小さい方法の代表として、コネクションマシンがある。コネクションマシンでは、一つのプロセッサ要素は1ビットの演算器と4Kビットメモリから構成されており、最も粒度を細かくしている。システム全体は単一命令で制御される単一命令複数データ流方式(SIMD(シングルインストラクション・マルチデータストリーム)方式)で行っている。1チップ(集積回路)に16個のプロセッサ要素を集積し、各チップは2進12-キューブ網で相互結合されている。全てのプロセッサ要素数は65536(2^{16})個である。コネクションマシンは人工知能用または画像処理用計算機として、非常に高性能なものとなっている。また、コネクションマシン上でニューラルネットワークをシミュレーションする報告がある。しかしながら、一般のニューラルモデルは複数ビットのデータ同士の乗算と加算の計算を繰り返すものが多く、1ビット演算器まで粒度を細かくしており、複数ビットの演算のためのオーバーヘッドがあり、これらの問題に対して、それほどの高速演算性能を期待できない。

【0003】また、ニューラルネットワークを高速に演算及び学習させる方法として、本願発明者らによる”高速学習

型ニューロWSIのシステム設計”、1990年10月25日電子情報通信学会、CPSY90-71、ICD90-127、及び特開平3-206549号公報等がある。これらは、複数のニューロンプロセッサをバスにより相互接続し、順次ニューロンプロセッサがそのニューロンの出力をバスにより、ブロードキャストすることにより、通信を行う。一般のニューラルモデルはニューロン間の結合が非常に多いので、このような単純な結合方法でも高速演算が可能である。通信のためのハードウェア量が少なく、時間的なオーバーヘッドが少ないため、非常にコンパクトにすることができる。また、ニューラルネットの欠陥に対する強さを利用して、複数の大面積集積回路のウェハスケール集積回路(WSI)により構成し、さらにコンパクトにしている。しかし、欠陥に弱い制御回路については大面積集積回路上に搭載していない。

【0004】また、ニューロン間の結合が疎な場合に効率よく、ブロードキャスト通信を行う方法として、USP 4,796,199が提案されている。ニューロンプロセッサを、ファミリー、グループ、コネクションといった階層的なまとまりに分け、各階層ごとにブロードキャスト通信を行うものである。ニューラルネットをモジュール化し、モジュールごとに効率的に学習させる場合モジュール間はモジュール内に比べ、通信量が少ないはずなので、そのような場合に高速に演算及び学習を行うことができる。しかし、その制御方法に関しては提案されていない。

【0005】一方、複数命令複数データ流方式(MIMD(マルチインストラクション・マルチデータストリーム)方式)として、Nキューブ2がある。Nキューブ2では、各プロセッサエレメントは独自に稼働し、各エレメント間はハイパーキューブトポロジーで結合されている。最大構成で8192(2^{13})個を設置面積4m²に実現している。

【0006】

【発明が解決しようとする課題】SIMDマシンでは、条件付き命令で若干の異なる動作を行うことができる。しかし、条件付き命令を複雑に分岐させれば、より複雑な動作が可能となるが、その場合、各プロセッサの命令デコードが複雑化していく。その最終的な形がMIMDマシンと考えることもできる。

【0007】SIMDマシン上で、モジュール化されたニューラルネットワークをシミュレーションする場合、各モジュールごとに大幅に異なるモデルを並列に実行することはできない。またMIMDマシンを利用する場合、同一のモジュール内のニューロンモデルを計算するにもかかわらず、それらのプロセッサエレメントは同一の命令セットが保持され、それらの制御回路は同一命令をデコードすることになり、外部から見た場合、非常に冗長な動作となる。また、1プロセッサエレメントのハードウェア量もSIMDマシンのものに比べ大きいため、

一定面積あたりのプロセッサ数は少なくなってしまう。

【0008】また、ニューラルネットの欠陥に対する強さを利用して、ニューロンモデルを計算する複数のニューロンプロセッサを1つの大面積集積回路上に搭載できるが、その制御回路は欠陥に弱いので、同一の大面積集積回路に搭載すると、歩留りが低下してしまうという問題があった。そのため、演算部と制御部は別の集積回路にする必要があった。

【0009】本発明の目的は、演算を行う機能ブロックを多数備えて、並列演算を行う情報処理装置をコンパクトに提供することにある。

【0010】本発明に他の目的は、演算を行う機能ブロックを多数備えて、並列演算を行う情報処理装置を更に複数個設け、超並列演算を行う情報処理システムを構築するにあたり、分散処理と論理演算を有効に行うことにある。

【0011】本発明の他の目的は、特に、ニューラルモデルの計算を高速に行うのに適した情報処理装置及びシステムを提供することにある。

【0012】

【課題を解決するための手段】本発明では、演算を行う機能ブロック(内積演算を高速に行えるものであって、例えば 10^4 個程度)とこれを制御する制御系を1つのモジュールとする。モジュール内ではSIMD方式により動作を行う。さらに複数のモジュール(例えば100個程度)を接続し通信を可能とする。複数のモジュールは、異なる命令セットであるMIMD方式により動作を行う。

【0013】また、ニューロンプロセッサ及びメモリなどの欠陥に対して強い部分は大面積集積回路上に集積し、欠陥に対して弱い制御部は、例えばシリコン・オン・シリコン(Si on Si)技術により、大面積集積回路上に制御部の集積回路を接続する。

【0014】

【作用】本発明によれば、モジュール化されたニューラルネットの学習及び動作を高速に行うことができる。各モジュールごと並列に学習が可能のため、高速な学習及び動作が可能である。プロセッサ数も、SIMD方式と同程度の実装密度とすることができ、モジュールごとに異なる命令セットにより動作させることが可能である。

【0015】また、本発明によれば、大規模の情報処理システムを非常にコンパクトに作ることができる。

【0016】

【実施例】実施例を用いて、具体的な構成について説明する。まず、その構成の概略について以下に説明する。

【0017】 10^4 個の機能ブロックである演算ブロック(内積演算を高速に行えるもの)を、例えば10cm角のウェハスケール集積回路(WSI)に集積し、かつ、スカラープロセッサ及びA/Dコンバータをシリコン・オン・シリコン技術で搭載し、1つのモジュールとす

る。アナログデジタル(A/D)コンバータにより、センサなどからのアナログ信号を入力することができる。このモジュール上には、スカラープロセッサのワーキングメモリも設け、これに 10^4 個の演算ブロックへの命令セットを記憶させておく。スカラープロセッサはワーキングメモリ及び各演算ブロックの出力、各演算ブロック内のローカルメモリをランダムにアクセスできる。また、各演算ブロックへの命令発行の管理も行う。各演算ブロックはモジュール内部のデータベースにより接続され、ブロードキャストを相互に高速に行える。 10^4 個のモジュール間は100ワードのモジュール間バスにより接続する。各モジュールには100個の 10^4 ワードの通信用バッファを持ち、モジュール間バスの各チャネルと一つずつ接続する。各モジュール内の内部データベースに出力された演算ブロックの出力は順次モジュール間バスを通し、全てのモジュール通信用バッファに書き込む。各モジュールでは、通信用バッファを読みだすことにより、外部モジュール上の演算ブロックの出力を知ることができる。書き込みタイミングはデータの転送側が発生することにより、モジュール間通信は非同期で行うことができる。

【0018】また、全モジュール内のワーキングメモリ、ローカルメモリをメモリ空間上に見ることができるスカラープロセッサを設ける。

【0019】以上の本発明の構成を図1～8を用いて以下に説明する。まず、図面を説明する。その後、図面を用いて動作について説明する。

【0020】図1に全体の構成図を示す。ここで、1～128はモジュール、130はモジュール間バス、140はモジュール入力バス、150はモジュール出力バス、141はグローバルプロセッサ、142はグローバルメモリを示す。グローバルメモリ142はグローバルプロセッサ141と接続し、グローバルプロセッサ141及びモジュール1～128は、モジュール入力バス140及びモジュール出力バス150を介して、モジュール間バス130により、相互に接続する。

【0021】図2はモジュール1～128のモジュールの構成を示す図である。200はモジュールでモジュール1～128と同じものである。201はローカルプロセッサ、202はワーキングメモリ203はアナログデジタル(A/D)変換器、204はニューロンプロセッサ、209はモジュール入力バッファを示す。210～328はモジュール入力バスで、それぞれグローバルプロセッサ141、モジュール1～128の信号を入力する。329はモジュール出力バスで、330は通信ユニットでモジュール入力バス211～328のそれぞれと接続する。400はモジュール内部バスを示す。ローカルプロセッサ201及びA/D変換器203は、シリコン(Si)基板上にシリコン基板をハンダバンプ等により直接接続する従来技術のシリコン・オン・シリコ

ン技術により、モジュール200上に接続することができる。ワーキングメモリ202とモジュール内部バス400はローカルプロセッサ201と接続する。モジュール200内のニューロンプロセッサ204及びA/D変換器203、通信ユニット330、モジュール入力バッファ209はモジュール内部バス400により相互接続する。モジュール入力バッファ209はモジュール入力バス210よりグローバルプロセッサ141からの信号を取り込み、モジュール内部バス400へ出力することができる。ニューロンプロセッサ204間は相互にモジュール内部バス400を使ってブロードキャストを行なうことにより通信することができる。211~328はモジュール入力バスでそれぞれモジュール1~128の出力信号を入力し、128個の通信ユニット330にそれぞれ取り込む。通信ユニット330に取り込まれた値は、ニューロンプロセッサ204と同様に、モジュール内部バス400を利用して、ブロードキャスト通信を行うことができる。モジュール内部バス400にブロードキャストされる値はモジュール出力バス329を介して、モジュール間バス130へ出力できる。

【0022】本発明をよりコンパクトにするために、ニューロンプロセッサ204及びワーキングメモリ202等の欠陥に対して強いブロックは1つの大面積集積回路601上に搭載する。そして、欠陥に対して弱い制御部、本図ではローカルプロセッサ201、A/D変換器203はシリコンオンシリコン技術で大面積集積回路601と接続する。

【0023】図3はニューロンプロセッサ204の構成を示す図である。また、ニューロンプロセッサ204の動作を制御する命令であるニューロン命令460も示す。ここで、401はモジュール内部入力バス、402はモジュール内部出力バスで、470は命令バスで、図2のモジュール内部バス400に対応する。403はAバス、404はBバス、405はCバス、411~413はフリップフロップ(FF)、421はワーキングレジスタ、422は乗算器、423はALU、424はシフト、425~426はレジスタファイル、427は重み値メモリ、428はトライステートバッファ、451~455は2-1または3-1のセクタである。Aバス403及びBバス404は乗算器422の入力信号である。ALU423はセクタ451とセクタ452を入力し、セクタ451はFF411またはAバス403を選択し、ALU423の一方の入力とする。セクタ452はBバス404または乗算器422を選択し、ALU423のもう一方の入力とする。セクタ453はALU423または乗算器422を選択し、Cバス405に出力する。FF411はCバス405の値を取り込むことができる。また、FF411はニューロン命令460によって、リセットすることができる。ワーキングレジスタ421及びレジスタファイル425~4

26、重み値メモリ427はCバス405の値を取り込むことができる。セクタ455はCバス405及びレジスタファイル425~426を選択し、トライステートバッファ428へ出力する。トライステートバッファ428の出力端子は、モジュール内部出力バス402へ接続し、その制御はニューロン命令460のニューロンプロセッサセレクト信号により行なう。これらの制御は、すべて命令バス470より送られるニューロン命令460により行う。

10 【0024】図4は図2の通信ユニット330の実施例の詳細を示す図である。ここで、501はバッファアレイ、502はバッファアレイ501の書き込みアドレスポインタ、503は1インクリメントで、アドレスポインタ502の値を1つつ進める。504は読みだし用のセクタで読みだしアドレス505により選択されたバッファアレイ501の値をモジュール内部出力バス402へ出力する。読みだしアドレス505は命令バス470より入力する。

20 【0025】図5は図2のローカルプロセッサ201から見たメモリ空間を示す図で、552はそのメモリ空間である。ここで、550はアドレス変換回路で、ローカルプロセッサ201からのアクセス要求に対し、物理的なアドレスに変換する。また、モジュール上のメモリの欠陥をさけ、リニア空間に見えるようにする。ローカルプロセッサ201からは、モジュール200上のワーキングメモリ202、レジスタファイル425~426、重み値メモリ427は同一メモリ空間上のデータとして見ることができる。また、ワーキングメモリ202上には、ニューロン命令460のセットを保持しておき、ローカルプロセッサ204はニューロン命令460を順次読みだし、ニューロンプロセッサ204へ送り、その動作を制御する。

30 【0026】図6は、モジュール200内において複数のニューロンプロセッサ204を並列演算させる場合を示す図である。本図に示すように、モジュール内部バス400はモジュール内部入力バス401とモジュール内部出力バス402と命令バス470で構成する。各ニューロンプロセッサ204は命令バス470により、ローカルプロセッサ201から送られるニューロン命令460を入力する。各ニューロンプロセッサ204はモジュール内部入力バス401とモジュール内部出力バス402と接続する。ニューロン命令460により、指定されたニューロンプロセッサ204はモジュール内部出力バス402にその出力を出力し、モジュール内部入力バス401を通して、各ニューロンプロセッサ204へ送る。各ニューロンプロセッサ204はニューロン命令460に従い、受け取ったデータを使って演算することができる。

40 【0027】図7は図1のグローバルプロセッサ141から見たメモリ空間を示す図で、602はそのメモリ空

間である。グローバルプロセッサ141からは、さらに図5のローカルプロセッサ201から見たメモリ空間552の各々が連続したメモリ空間に見ることができる。図5及び図6の構成とすることにより、ローカルプロセッサ201及びグローバルプロセッサ141は非常に大きなメモリ空間を持つスカラプロセッサとして動作することが可能であり、かつ、内積演算などのニューロ的な動作も、そのメモリ空間上で行うことができる。

【0028】これまでは、このような並列計算機はホストコンピュータと接続され、ホストコンピュータ上でデータを加工したり、作成したりした後、並列計算機側へデータ及びプログラムをロードして、行わなければならない。本発明では、全く同一メモリ空間上でスカラ*

$$\tau du_i/dt = -u_i + \sum w_{ij} x_j + I_i \quad (1)$$

$$x_i = f(u_i) \quad (2)$$

で表せる。ここで、Iはいわゆる定電流源であるが、一般には、常に、その最大値を出力するニューロンを設け、それに対して重み付けしてシナプス結合すること ※

$$f(u_i) = 1 / (1 - \exp(-u_i/T)) \quad (3)$$

などの、飽和関数を用いられる。

【0033】1～3式をディジタル表現で計算する場合、1式を時間刻み幅 Δt で差分化する。すべてのニューロンの内部エネルギー及び出力をそれぞれベクトル ★

$$u_{t+1} = u_t - \Delta t / \tau (W x_t - u_t) \quad (4)$$

$$x_{t+1} = f(u_{t+1}) \quad (5)$$

の計算を各時間ごとに行なえばよい。4～5式をモジュール200上で行なう方法を以下に示す。

【0034】図10は4式の $W x_t$ を行なう場合のニューロンプロセッサ204の各演算回路及びメモリ、FFの接続方法を示した図である。これらの接続はニューロン命令460により設定することができる。各ニューロンプロセッサ204を各ニューロンに対応させる。各ニューロンプロセッサ204では、乗算器422はモジュール内部入力バス401と重み値メモリ427を入力する。ALU422は乗算器422とFF411を入力し、その加算結果をFF411に書き込む。ローカルプロセッサ201は順次、ニューロンプロセッサ204を選択し、選択されたニューロンプロセッサ204はその出力 x_t をモジュール内部出力バス402に出力する。☆

$$x_{t+1} = \alpha_0 + \alpha_1 u_{t+1} + \alpha_3 u_{t+1}^3 + \alpha_5 u_{t+1}^5 + \alpha_7 u_{t+1}^7 + \dots \quad (6)$$

ここで $\alpha_0 \sim \alpha_7$ は関数 f によってきめる定数、などによって計算することができる。この計算も複数のニューロン命令460を行なうことにより可能である。

【0035】以上に示したように、ニューロンプロセッサ204の動作は、ニューロン命令460により、自由に決めることができるため、任意のモデルを計算することができる。

【0036】図11、12は2つのモジュール上で相互結合型ニューラルネットワークを実現する場合を示した図である。例えば、ニューロン11、12をモジュール 50

*一処理及び並列処理を行うことができる。

【0029】図8はモジュール1～128間の通信方法を示す図である。

【0030】まず、各モジュール200でニューラルモデルを演算させる場合について、説明する。

【0031】図9は相互結合型ニューラルネットワークを1枚のモジュール200上で実現する場合を示した図で、以下にそのふるまい及び一般的なモデル式を説明する。

【0032】ニューロン i の出力を x_i 、内部エネルギーを u_i 、ニューロン j に対する重み値を w_{ij} とすると、各ニューロンは状態方程式、

※で、省くことができる。また、2式の f は非線形関数を用いられ、一般には、シグモイド関数、

20★ u 、 x で、また、すべての重み値を行列 W で表し、時刻 t でのベクトルを u_t 、 x_t 、時刻 $t+1$ でのベクトルを u_{t+1} 、 x_{t+1} とすると、

☆予め、ニューロンプロセッサ204は固有のアドレスを割りふっており、ローカルプロセッサはそのアドレスを発生し、アドレスをデコードすることにより、上記の制御を行なうことができる。図2に示すように、各ニューロンプロセッサ204は x_t をレジスタファイル425に保持し、セクタ455、トライステートバッファ428を通して、出力することができる。また、4式における Δt 及び τ 、 u_t をワーキングレジスタ421またはレジスタファイル425、重み値メモリ427に保持しておき、ALU423、乗算器422、シフト424、ワーキングレジスタ421を仕様して、4式を計算することができる。5式の非線形変換は、例えば、チェビシェフ近似、

1のニューロンプロセッサ204に割りふり、ニューロン21、22をモジュール2に割りふる。

【0037】各モジュールは個別に、順次ニューロンプロセッサ204がモジュール内部バス400を用いて、ブロードキャストする。それと同時に、モジュール1のモジュール内部バス400に出力された値は、モジュール出力バス329を通して、モジュール間バス130に出力し、各モジュール200の通信ユニット330へ送られる。各モジュール200では、自分以外の各モジュール200に対応する通信ユニット330を持ってお

り、それに対応するモジュール間バス130に接続する。各通信ユニット330への書き込みは送り手側のタイミングで行う。図12では、モジュール1では通信ユニット330を読みだせば、モジュール2上のニューロン21、22の出力を読むことができる。同様に、モジュール1上では、モジュール2上のニューロン11、12の出力を通信ユニット330を通して、読むことができる。図4に示すように、通信ユニット330は、順次モジュール入力バス140により送られる値を書き込みアドレスポインタ502の差すバッファに取り込む。書き込むと同時にアドレスポインタ502を1インクリメント503により1つ進める。また、バッファアレイ501の各出力はセクタ504を通して、読みだしアドレス505により指定してモジュール内部入力バスへ読みだすことができる。バッファアレイ501への書き込むタイミングはモジュール入力バス140によって、データと同時に送る。また、図1、図8に示すように、すべてのモジュール200とモジュール間バス130を同様に接続することで、全てのニューロンプロセッサ204間の通信を行うことができる。また、各通信ユニット330への書き込みは、送り手側のタイミングで行うので、各モジュールは独自のタイミングと独自の命令セットに従い、動作することができる。

【0038】また、モジュール間バス130は配線長が長く、高負荷となる可能性がある。その場合、モジュール200の内部は高速に動作するにも係らず、モジュール間バス130はそのスピードに追従できない可能性がある。しかし、その場合でも、本発明はモジュール間の通信は非同期で行うことができるので、例えば、モジュール200の内部は100MHzで動作し、モジュール間バス130は50MHzで動作するということも可能である。そのときのニューロン間の通信は、時間方向に間引いて通信すればよい。

【0039】同様に複数のモジュール間の通信も可能であり、通信ユニット330のバッファアレイ501を各モジュールのニューロンプロセッサ204の個数だけ用意すれば、全ニューロンの完全結合が可能である。また、通常は、モジュール間はモジュール内に比べ、結合が少ないので、それよりは少ないバッファ数を用意してもよい。

【0040】図13に本発明の利用例を示すが、本図に示すように、各モジュール1～128を個別のアルゴリズム及び個別のデータにより、学習及び自己組織化が可能であり、また、それらは、並列に動作させることが可能である。また、図11、12に示したように、複数のモジュール200を用いて、同一のモデルを動作させることも可能である。図13では、モジュール1～2及び127～128をバックプロパゲーション(Back Propagation)により学習を行い、モジュール3を学習ベクトル量子化(Learning Vector Quantization)により学習を

行い、モジュール4を競合学習により学習を行い、モジュール125～126はホップフィールド型ネットワークとして使用する場合は示している。例えば、ホップフィールド型ネットワークにより、ノイズ除去及びエッジ検出などの初期視覚を行い、その結果をバックプロパゲーションにより学習を行う階層型ネットワークに入力し、パターン認識を行うことが可能である。また、ベクトル量子化モデルにより、入力した文字画像の文字認識を行い、その結果を階層型ネットワークにより音韻データを出力し、さらに、別のネットワークにより音声合成することが可能である。また、本図に示していないモジュールにおいても、同様に、利用することができる。

【0041】図14の本発明の利用例には、機能的な表現で動作を表わしている。ロボットの頭脳として、利用する場合を示している。10°程度のプロセッサ数では、到底ロボットの頭脳をすべて作ることはできないが、その一部の動作を行うことができる。例えば、モジュール1により、入力した画像の特徴検出等を行い、モジュール3の画像の記憶と比較する。また、入力した音声信号をモジュール2、4により認識する。画像認識結果、音声認識結果または両結果の総合した結果に対して、モジュール127が動作パターンを発生する。モジュール125には動作パターンを記憶させておく。モジュール128を使って、動作パターンから実際のロボットの関節を動かす信号へ変換する。また、ロボットが行動した結果をモジュール126により評価し、よりスムーズな動作パターンを学習して行くことができる。

【0042】また、本発明によれば、10°以上のプロセッサ数からなるシステムも同様に可能である。図2に示すように、各モジュールには、ローカルプロセッサ201を搭載するので、これにより、従来のAI(人工知能)技術にある記号処理や定性推論を行うことも可能である。また、ニューロンプロセッサを利用して、ファジー理論のメンバーシップ関数を決定し、ローカルプロセッサ201またはグローバルプロセッサ141により、ファジー理論による判断決定を行うこともできる。あるモジュールでは、ニューロンモデルによる情報処理を行い、別のモジュールでは、ファジー理論による情報処理を並列に行うことも可能である。

【0043】次に本発明の実装方法、組立て方法及び冷却方法について、図15～18を用いて説明する。

【0044】図15はモジュール200のシリコン・オン・シリコンを説明する図である。601は大面積集積回路、602は集積回路を示す。603は大面積集積回路601と集積回路602を接続するハンダのバンプである。605はバンプ603を接続するパッドである。本図に示すように、集積回路602と大面積集積回路601の接続する信号線及び電源給電のパッドを向い合わせバンプを介して、固定及び信号線の接続を行うことができる。

10

20

30

40

50

【0045】図16はボードと図15の大面積集積回路601の接続方法を説明する図である。605はパッド、606はコネクタで別のボードと接続するためである。610は接続用ピンで別のボード606と接続し、スタックすることができる。図15で示したシリコンとシリコンの接続方法と同様に、ボード上のパッド605と大面積集積回路601上のパッド605を向い合わせ、バンプ603により接続することができる。

【0046】図17はボード間の接続方法及び組立て方法を示す図である。本図で、800は接続ボードで、801はボードを示す。805、820はコネクタを示す。ボード801上には、グローバルプロセッサ141とグローバルメモリ142を載せる。本図では、4枚のボード606をスタック状に接続したものを示した図であるが、さらに、多くのボードも同様に接続することができる。ボード606は接続用ピン610で接続し、さらに、4辺に接続ボード800を接続する。ボード606のコネクタ607と接続ボード800のコネクタ805をそれぞれ対応させて接続する。さらに、接続ボード800のコネクタ820とボード801のコネクタ805をそれぞれ対応させて接続する。

【0047】次に各集積回路の冷却方法について説明する。図17の構成では、各ボード間の信号線を4辺を使ってボード間接続を行うので、非常に多くのボード間接続を行うことができる。しかし、図17からわかるように、内部が密閉されるため、動作中高温となる可能性がある。その場合は、図18に示す方法により、集積回路を冷却することができる。本図で、901は冷却水、910は冷却板を示す。ボード606間を接続する接続用ピン610を冷却水901の通路とする。各ボード60*30

$$Ax = b$$

を解くことになる。ここで、Aは $n \times n$ の行列、bは n 次ベクトル、xは求めるべき未知数である n 次ベクトルである。nは求める系の次元である。6式の方程式の解法として、いくつかの方法があるが、ここでは、線形緩和※

$$A = L + D + U$$

ある初期値 x_0 により以下の反復式、

$$x_{k+1} = 1/D (b - (L + U) x_k)$$

を反復して収束すれば、その値が解となる。8式を本発明に解く場合、各ニューロンプロセッサ204にベクトルxの各要素を計算させる。各ニューロンプロセッサ204の重み値メモリ427には、行列Aの各行要素を格納する。対角要素だけは、予め逆数を求めておき、ワーキングレジスタ421に格納しておいてもよい。また、ベクトルbの要素もワーキングレジスタ421または重み値メモリ427のいずれかに格納しておけばよい。前回の反復結果 x_k はレジスタファイル425に格納しておく。ローカルプロセッサ201は順次、各ニューロンプロセッサ204を指定し、前回の反復結果 x_k をモジュール内部バス400にブロードキャストしていく、各

*6では、接続用ピン610から冷却水901を大面積回路601に接続する冷却板910に引き込む。使用された冷却水901は反対側の接続用ピン610に戻す。このようにして、密閉された内部の集積回路を冷却させることができる。

【0048】次に図19を用いて、大面積集積回路を高速に動作させるために、必要なクロックのスキューを低減するための、クロック給電方法を説明する。図2に示すような大面積の集積回路で同期した動作を行うには、信号遅延のため大きなクロックスキューが生じ、高速動作させるのは難しい。そのような場合、図19に示すように、外部から入力したクロック信号をまず、大面積集積回路中央へ引き込む、さらに、中心から端辺までの半分の位置、さらに半分の位置とクロック信号線引き回し、大面積集積回路内のどの位置でも、等長に近く配線することができる。そのため、大面積集積回路内のクロックスキューを低減することができる。

【0049】以上の接続方法、組立て方法及び冷却方法により、図1の構成を実現することができる。

【0050】以上は本発明上でニューラルネットモデルを高速に演算させる場合について説明した。次に、他の情報処理を行う場合の実施例を示す。

【0051】一般的に、物理現象は、微分方程式により表わされ、それを解くことにより、計算機上でシミュレーションを行っている場合が多い。これらの方程式は、積分公式や差分方程式化により、各時間ステップごとの非線形方程式に表わし、さらに、ニュートン・ラプソン法等により非線形方程式の解を求め、各時間ごとの状態を決定する。また、ニュートン・ラプソン法の各反復ステップでは、線形化された連立方程式、

$$(6)$$

※和法による場合について説明する。線形緩和法の一つとして、ヤコビ法がある。ヤコビ法では、行列Aを下三角成分Lと対角成分Dと上三角成分Uに分ける。

$$[0052]$$

$$(7)$$

$$(8)$$

ニューロンプロセッサ204では、8式の $(L + U) x_k$ に対応する行列Aの要素を重み値メモリ427から読みだし、乗算器422により乗算を行い、ALU423とFF411を使用して、順次累積加算することにより計算することができる。一通りブロードキャストが終了すれば、8式の $(L + U) x_k$ の計算結果がわかり、さらに、8式の他の計算を行い、最後に対角成分の逆数を掛けることにより、この回の反復結果 x_{k+1} を決めることができる。その結果は、レジスタファイル425に書き込み、次の反復に進む。ある程度反復回数が進んだところで、収束判定を行い、収束していれば、6式が解けたことになる。また、複数のモジュール200を用い

て、大規模な方程式に対しても、ニューラルネットモデル同様に、計算することができる。

【0053】また、ニューラルネットモデル1～3式は、まさに連立非線形微分方程式であり、前述の計算方法は前進オイラー法により積分しているものである。このように、本発明は、ニューラルネットモデルを高速に計算するだけでなく、一般の数値シミュレーション等にも広く適応できるものである。その他、画像処理等にも、適応することができる。例えば、動画像を入力し、各フレームを各モジュールに順次ロードする。各モジュールでは、各フレームのノイズリダクション及びエッジ検出等を行う。さらに、過去のフレームの情報を通信ユニット330により読むことができるので、オプティカルフローなど、動画像の情報処理に必要な情報を計算することができる。

【0054】次に、本発明をワークステーションに利用した例を図20に示す。本図で、1100はリスクプロセッサ、1101はI/Oプロセッサ、1102はディスクコントローラ、1103はグラフィックコントローラを示す。大面積集積回路601上にリスクプロセッサ1100及びI/Oプロセッサ1101等を図15に示したシリコンオンシリコンで接着する。リスクプロセッサ1100は大面積集積回路601のワーキングメモリ202にオペレーティングシステム等のプログラムを保持し、その命令に従い、動作する。また、ニューロンプロセッサ204の出力及びローカルメモリもワーキングメモリ202と同様のメモリ空間として見ることができる。このような構成としたワークステーションでは、上記のニューラルネットの計算はもとより、物理シミュレーション等の数値演算などニューロンプロセッサ204の演算器を使って、高速に行うことができる。

【0055】

【発明の効果】本発明によれば、非常に多くのニューロンから構成されるニューロコンピュータを実現できる。また、アプリケーションの開発もモジュール化して、システムを構築して行くことができるので、非常に効率よく行うことができる。また、ニューロコンピュータのみならず、物理シミュレーションも高速に計算することができる。ワークステーションのエンジンに利用すれば、ワークステーションのアプリケーションを広げることができる。

【図面の簡単な説明】

【図1】本発明の全体構成を示す図。

【図2】本発明の1モジュールをウェハスケール集積回路で実現した例を示す図。

【図3】ニューロンプロセッサの構成例を示した図。

【図4】モジュール間通信を行うための通信ユニットを示した図。

【図5】ローカルプロセッサのメモリ空間を説明する図。

【図6】ブロードキャストアーキテクチャを説明する図。

【図7】グローバルプロセッサのメモリ空間を説明する図。

【図8】モジュール間の通信方法を示す図。

【図9】相互結合型ニューラルネットワークを示した図。

【図10】相互結合型ニューラルネットワークのブロードキャストアーキテクチャ上での演算方法を示す図。

【図11】2つのモジュール上で相互結合型ニューラルネットワークを演算される例を示した図。

【図12】2つのモジュール上で相互結合型ニューラルネットワークを演算される例を示した図。

【図13】本発明の利用例を示す図。

【図14】本発明をロボットの頭脳の一部に利用した例を示す図。

【図15】シリコン・オン・シリコン技術を説明する図。

【図16】シリコン・オン・ボード技術を説明する図。

【図17】ボード間接続を説明する図。

【図18】冷却方法を説明する図。

【図19】クロック等長配線を説明する図。

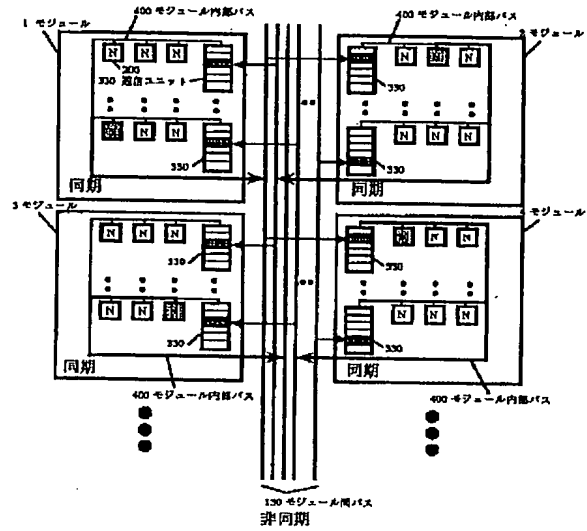
【図20】本発明をワークステーションに利用した例を示す図。

【符号の説明】

1～128・・・モジュール、130・・・モジュール間バス、140・・・モジュール入力バス、141・・・グローバルプロセッサ、142・・・グローバルメモリ、150・・・モジュール出力バス、200・・・モジュール、201・・・ローカルプロセッサ、202・・・ワーキングメモリ、203・・・A/D（アナログデジタル）変換器、204・・・ニューロンプロセッサ、209・・・モジュール入力バッファ、210～328・・・モジュール入力バス、329・・・モジュール出力バス、330・・・通信ユニット、400・・・モジュール内部バス、601・・・大面積集積回路。

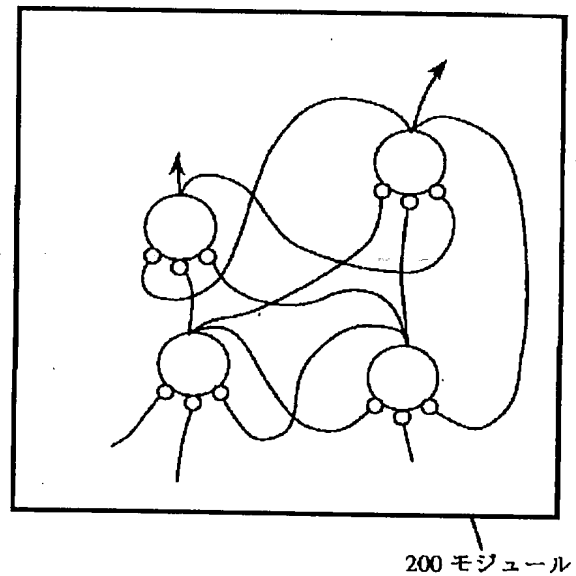
【图 8】

图 8



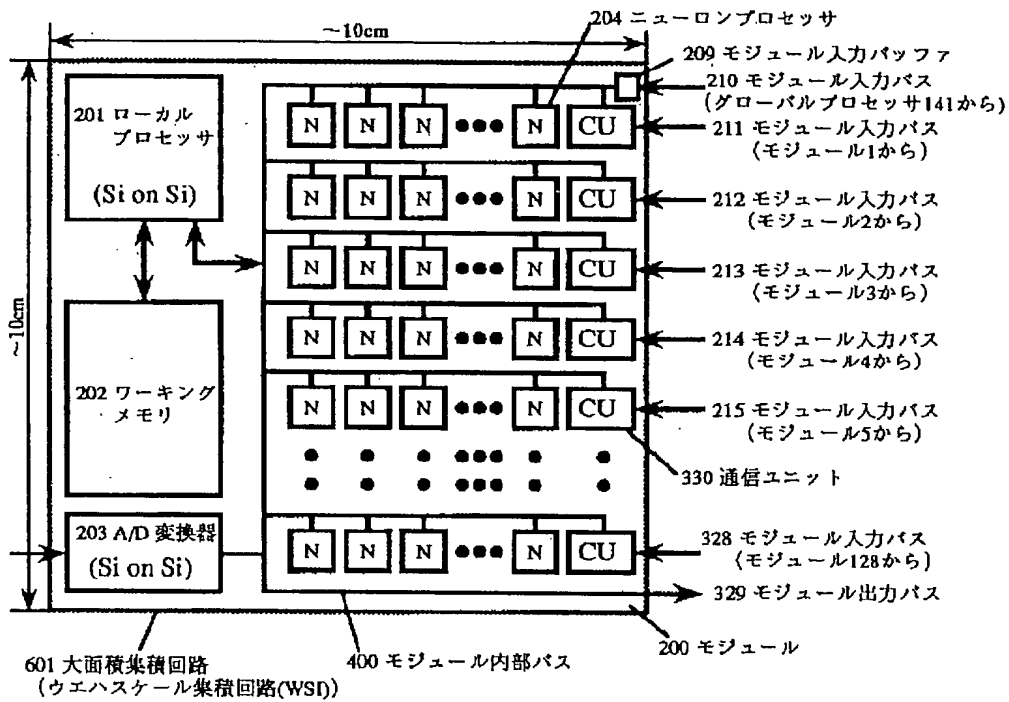
【図 9】

图9



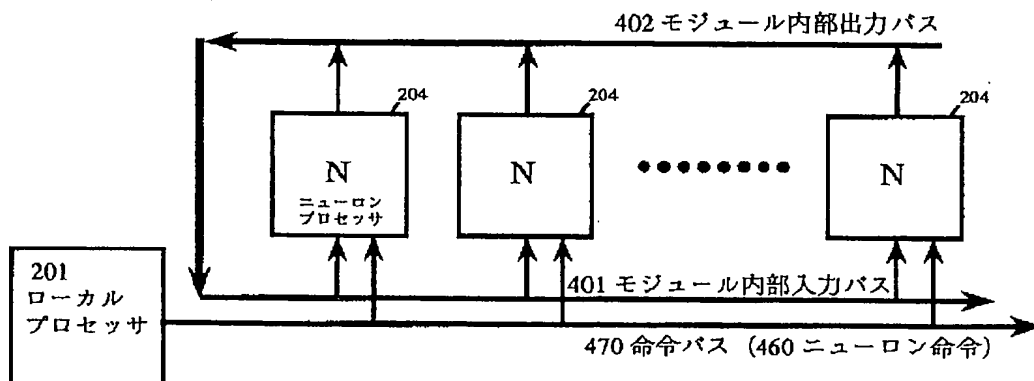
【図 2】

図 2

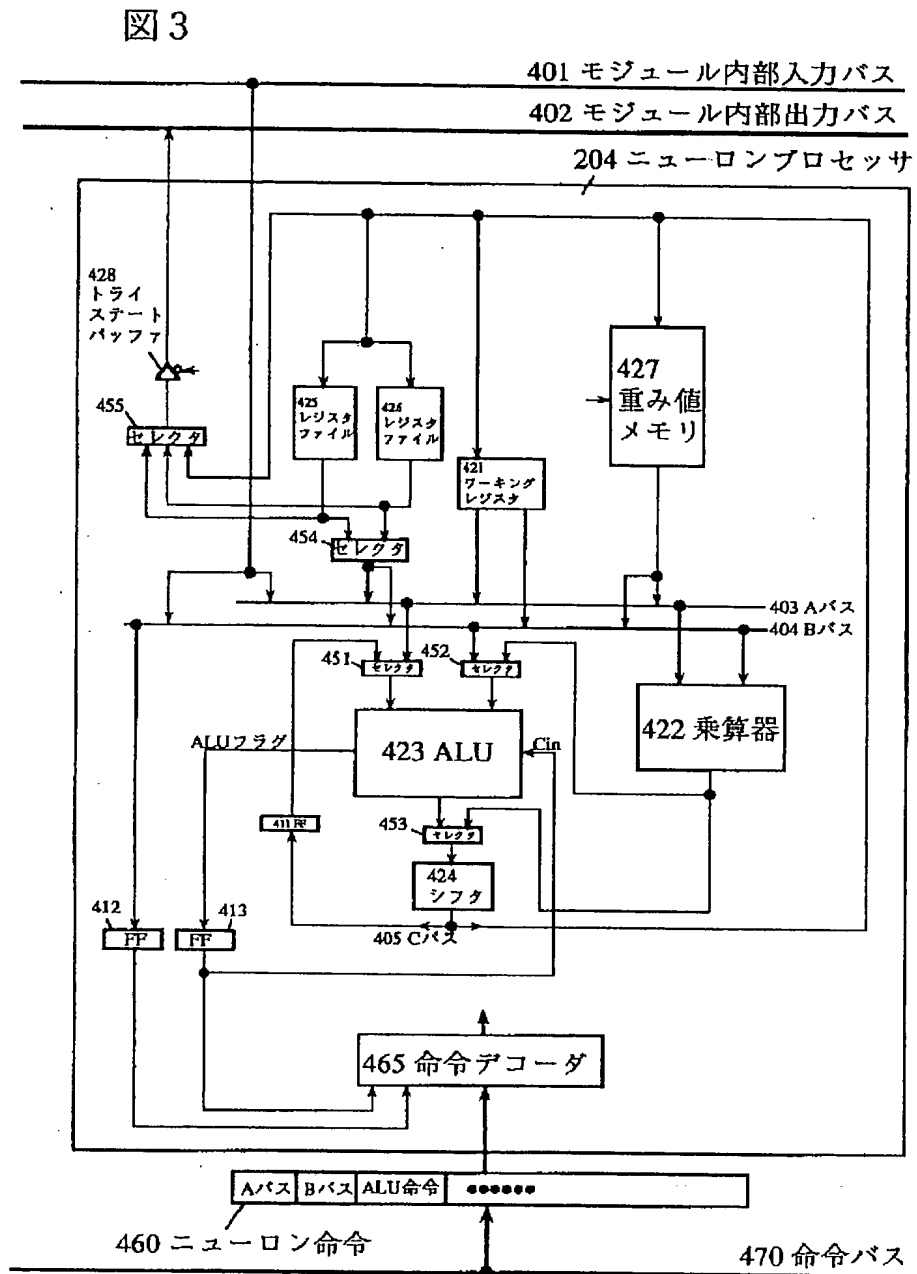


【図 6】

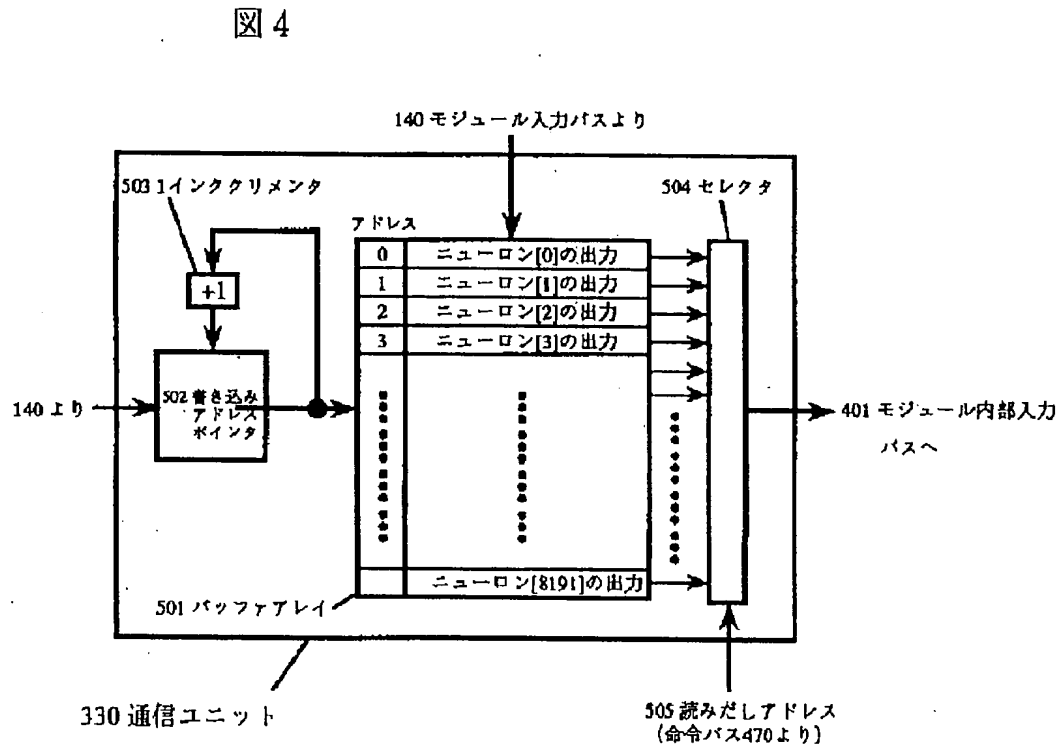
図 6



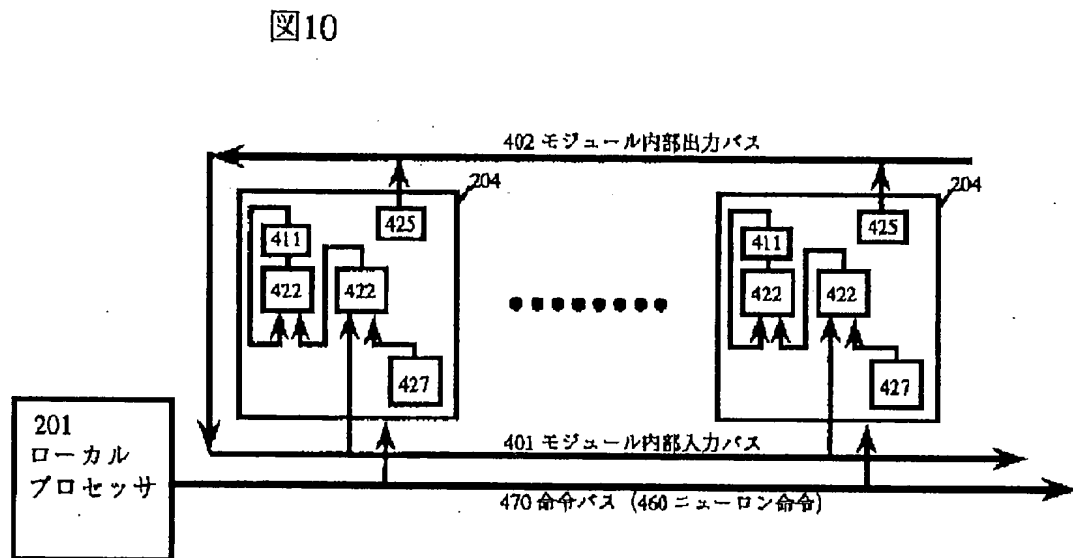
【図3】



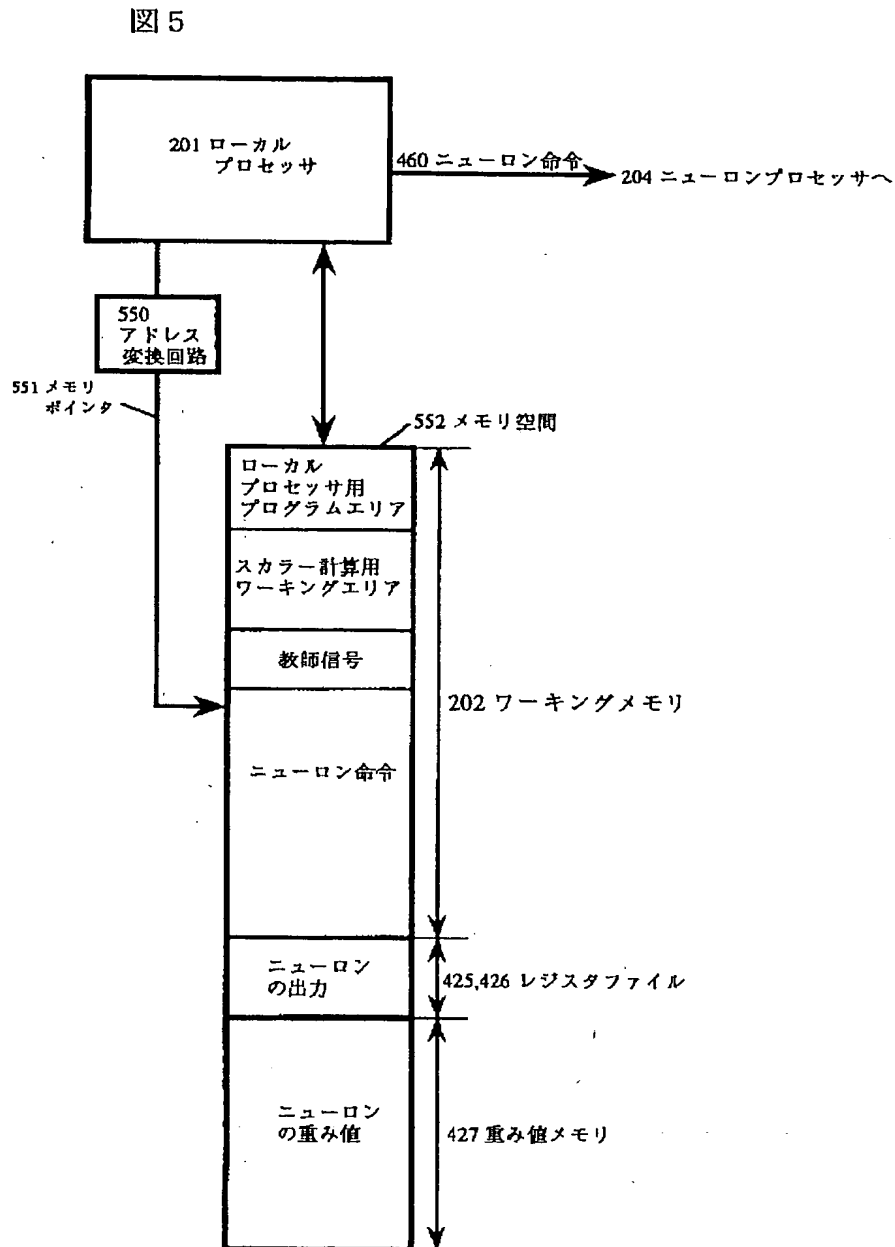
【図4】



【図10】

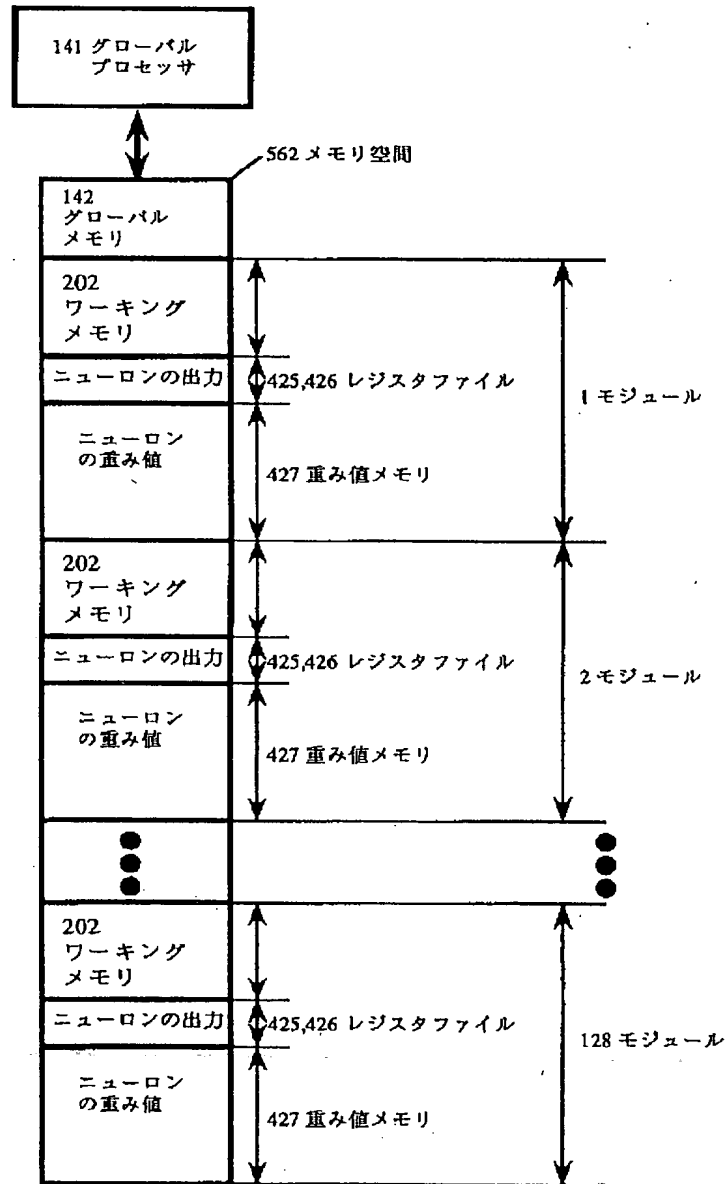


【図5】



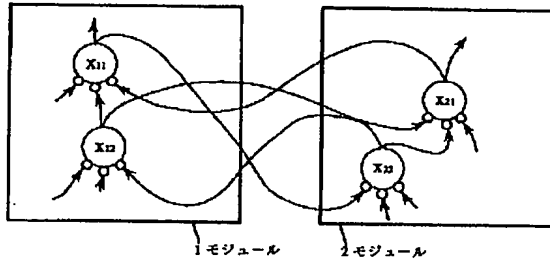
【図7】

図7



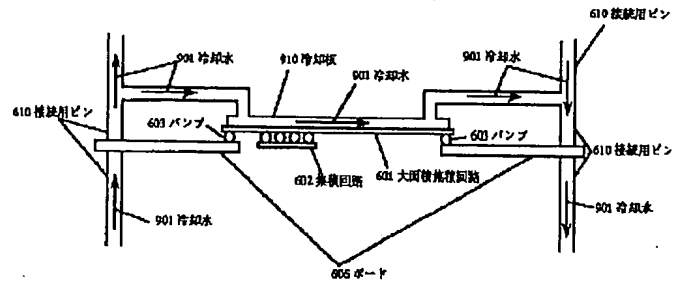
【図11】

図11



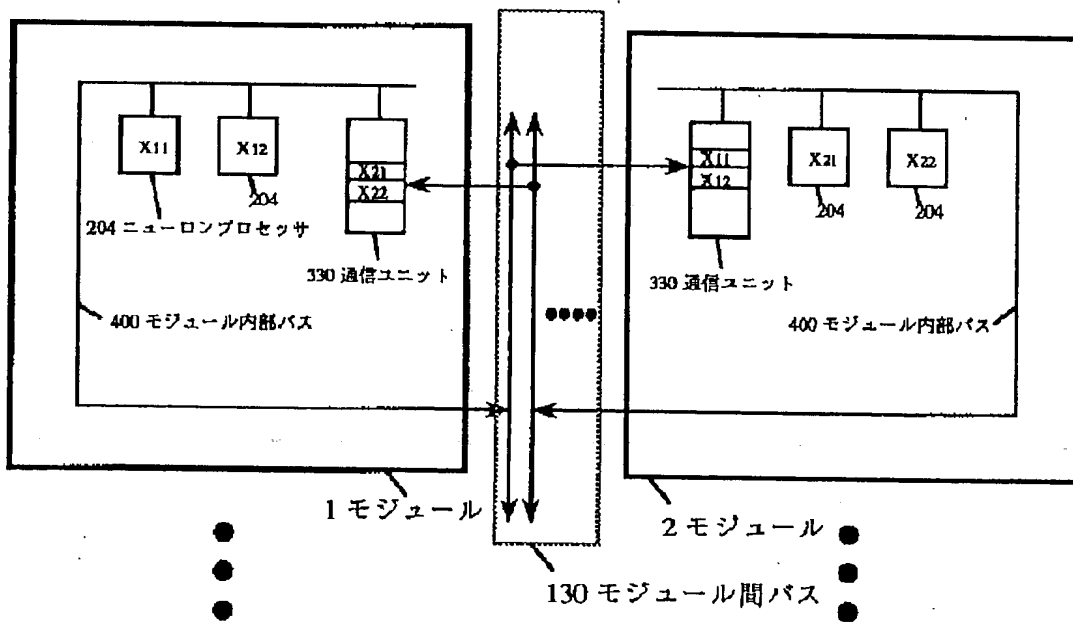
【図18】

図18

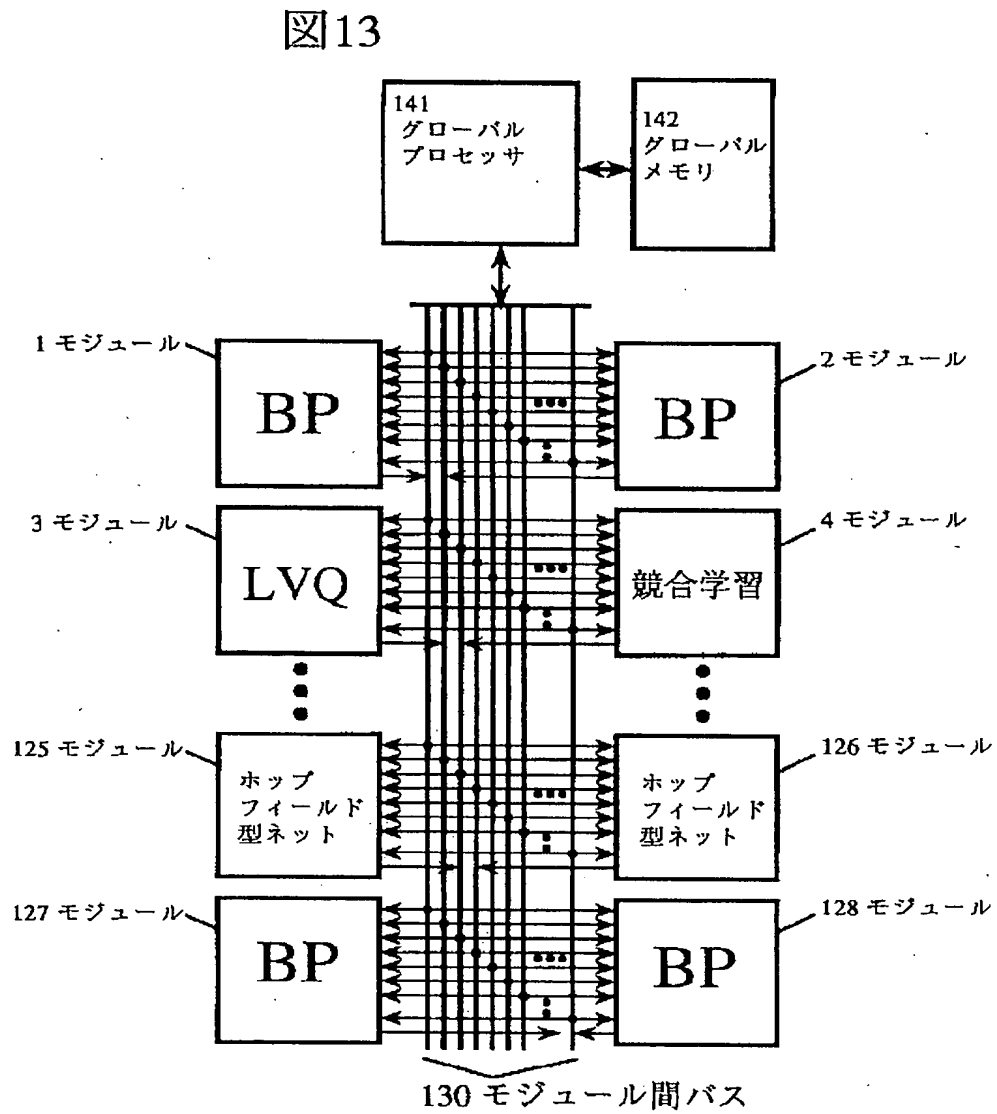


【図12】

図12

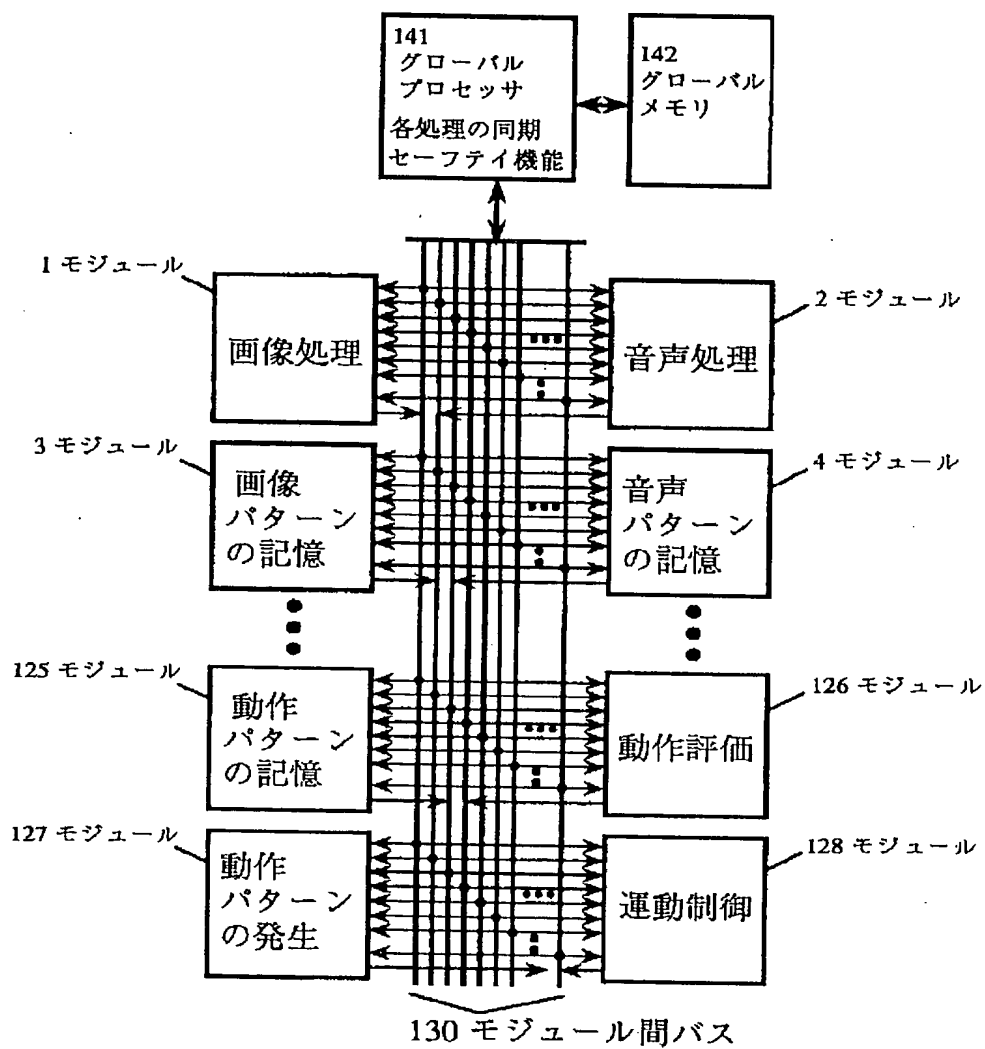


【図13】



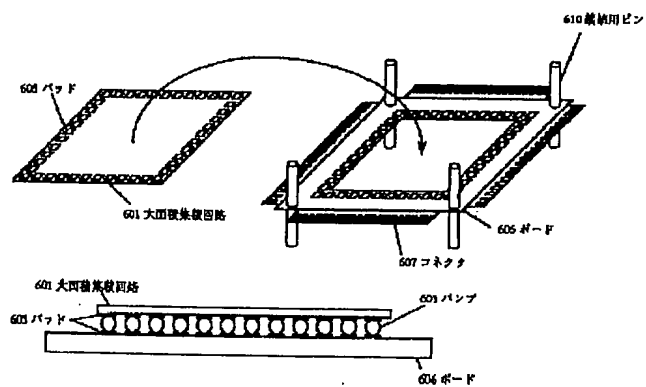
【図14】

図14



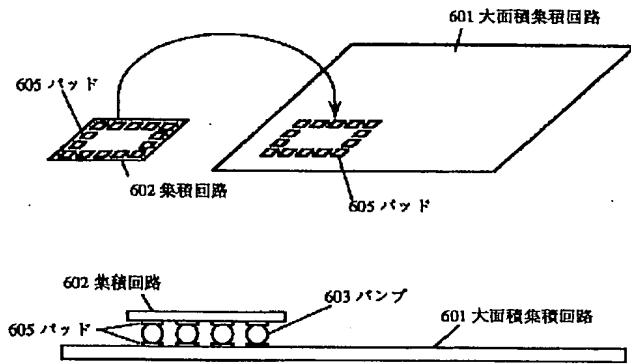
【図16】

図16



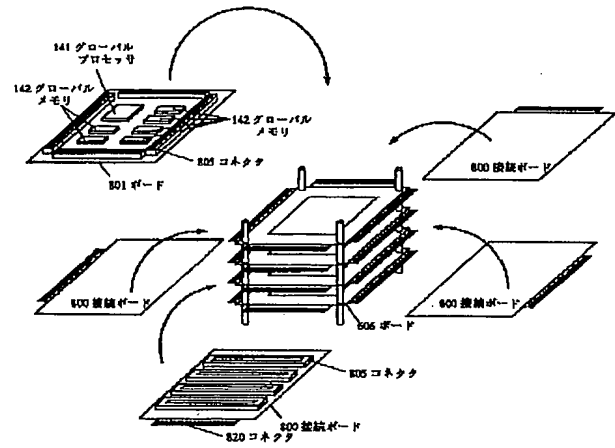
【図15】

図15



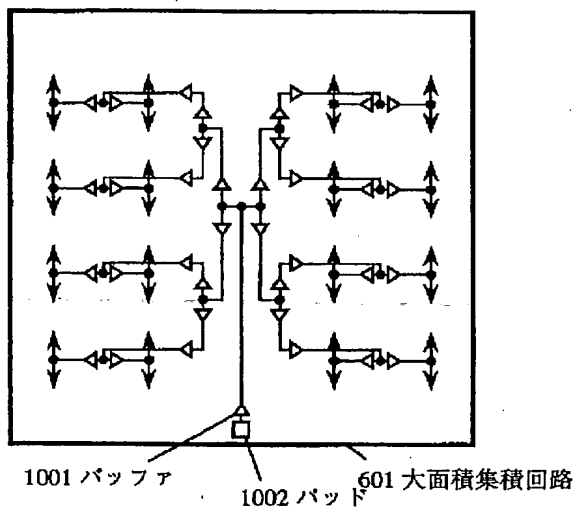
【図17】

図17



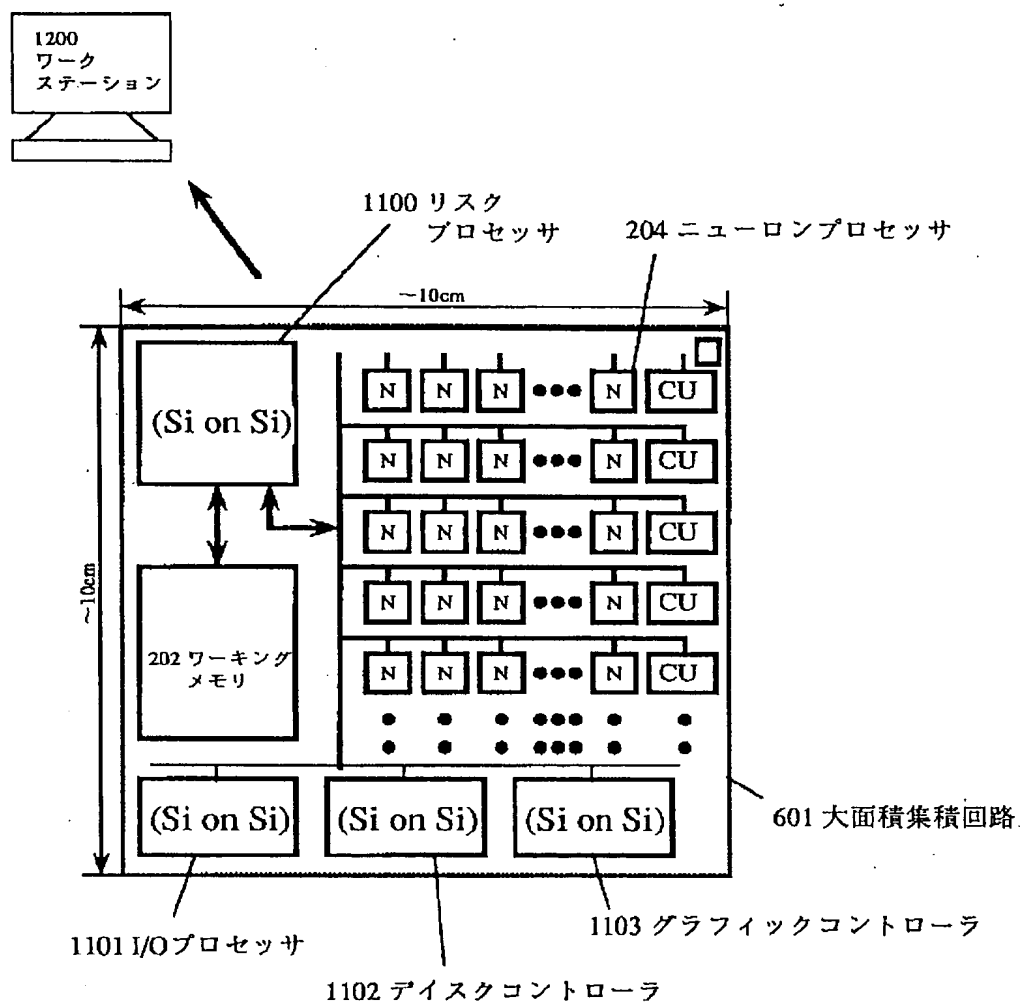
【図19】

図19



【図20】

図20



フロントページの続き

(72)発明者 山田 稔
 東京都国分寺市東恋ヶ窪1丁目280番地
 株式会社日立製作所中央研究所内

(72)発明者 坂口 隆宏
 東京都小平市上水本町5丁目20番1号 日
 立超エル・エス・アイ・エンジニアリング
 株式会社内
 (72)発明者 橋本 雅
 東京都小平市上水本町5丁目20番1号 日
 立超エル・エス・アイ・エンジニアリング
 株式会社内